IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156107

Scalable Machine Learning Pipeline for U.S. Visa Approval Prediction Using AWS

Gowthama T[1], Sandarsh Gowda M M[2]

Department of Master of Computer Applications, Bangalore Institute of Technology, Bengaluru, India[1]

Assistant Professor, Department of Master of Computer Applications, Bangalore Institute of Technology, Bengaluru, India[2]

Abstract: The U.S. visa approval process involves complex, high-volume decisions shaped by applicant and employer attributes. This essay offers a comprehensive Machine Learning Operations pipeline for automated visa approval prediction using the Easy Visa dataset (25,480 records, 12 features). The system integrates automated data validation, schema enforcement, class imbalance handling, and feature engineering within a reproducible pipeline deployed on Amazon Web Services (AWS) . Several classifiers were assessed under the same conditions, the KNN classifier achieved 96.83% accuracy and the highest F1-score. The pipeline incorporates continuous monitoring, drift detection, and automated model promotion, transforming a static classifier into a sustainable decision-support framework. Results confirm the viability of end-to-end Machine Learning Operations integration for administrative classification tasks.

Keywords: MLOps, visa prediction, Amazon Web Services, data drift detection, binary classification, cloud deployment, K-Nearest Neighbors.

I. INTRODUCTION

The adjudication of U.S. visa applications is an inherently data-intensive process shaped by employer characteristics, wage levels, occupational requirements, and prevailing economic conditions. With application volumes continuing to rise, traditional rule-based screening mechanisms face limitations in scalability, consistency, and timeliness. Automated machine learning systems offer a compelling alternative, yet most existing work treats visa approval prediction as a standalone classification task without addressing the full operational lifecycle.

Machine Learning Operations (MLOps) bridges the gap between experimental model development and reliable deployment by automating training, evaluation, monitoring, and retraining workflows. When combined with cloud infrastructure such as Amazon Web Services (AWS), Machine Learning Operations enables reproducible and auditable pipelines suitable for high-stakes administrative applications.

This paper presents a complete Machine Learning Operations pipeline for U.S. visa approval prediction. The system ingests historical visa application data, performs automated preprocessing and feature engineering, trains and evaluates multiple classifiers, and deploys the best-performing model to a cloud-based inference endpoint with continuous monitoring and drift detection. The K-Nearest Neighbors classifier achieved 96.83% accuracy and was selected as the production model. The primary contributions are: (i) an end-to-end Machine Learning Operations framework tailored for administrative classification; (ii) automated schema validation and drift reporting integrated into the pipeline; (iii) a containerized AWS deployment with real-time prediction through an API-driven interface.

II. LITERATURE REVIEW

Automated visa approval prediction has attracted increasing research interest. Early work by Ahmad et al. [1] applied logistic regression and decision tree classifiers to immigration datasets, demonstrating that employer attributes are strong predictors of approval outcomes, but without any deployment consideration. Sharma and Gupta [2] explored Random forest- based ensemble techniques, reporting improved accuracy over single classifiers, though model version control and retraining pipelines were absent.

Vegesana [8] demonstrated the superiority of ensemble methods over single classifiers but did not incorporate automated retraining. From an infrastructure perspective, Konda [11] analyzed cloud-based AI/ML deployment strategies, underscoring containerization benefits without addressing domain-specific visa governance requirements. AWS resources [12], [13] provide architectural guidance on lifecycle management and scalable deployment using S3, EC2, and SageMaker, yet remain generic and lack application-specific validation in immigration analytics.

Advancements in explainable AI highlighted by Wang et al. [15] stress transparency in decision-support systems, yet explainability remains underutilized in production-grade visa prediction systems. Overall, significant gaps persist in

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156107

integrating end-to-end Machine Learning Operations pipelines, drift detection, model governance, and cloud-based deployment within a unified framework. Table I summarizes key related works and their limitations

TABLE I

Comparison of Related Works in Visa Approval Prediction

Author	Method	Dataset	Acc.	Limitation
Ahmad et al. [1]	Logistic Reg., Decision Tree	Immigr. Records	~87%	No deployment; static
Sharma & Gupta [2]	Random Forest, SVM	H-1B Dataset	~90%	No retraining/drift
Vegesana [8]	Ensemble Methods	Kaggle Visa	~93%	No version control
Konda [11]	Cloud AI/ML Arch.	General	N/A	Not domain-specific
Proposed	KNN, XGBoost, RF + MLOps	EasyVisa	96.83%	Full Machine Learning Operations lifecycle

In order to fill these deficiencies, the proposed approach integrates a complete MLOps lifecycle — data ingestion, validation, preprocessing, model training, evaluation, deployment, and continuous monitoring — within a unified cloud-based framework.

III. PROBLEM DEFINITION AND RESEARCH MOTIVATION

The process of evaluating U.S. visa applications involves analyzing a large number of factors related to applicants, employers, and employment conditions. These decisions are influenced by historical approval patterns, regulatory constraints, and economic indicators, making the task inherently complex. As application volumes increase, traditional manual assessment and rule-based screening face limitations in scalability, consistency, and timely decision support.

Most existing solutions treat visa approval prediction as a standalone classification task focusing primarily on model accuracy in isolated experimental settings. Such approaches do not adequately address practical deployment challenges, including data schema consistency, temporal drift, and systematic model lifecycle management. The research motivation is therefore to design a complete, automated Machine Learning Operations pipeline that achieves excellent forecast accuracy while guaranteeing reproducibility, auditability, and resilience in production.

IV. DATASET DESCRIPTION AND EXPLORATORY DATA ANALYSIS

The study utilizes the EasyVisa dataset sourced from the public Kaggle repository [14]. The dataset comprises 25,480 records and 12 features encompassing employer characteristics, job attributes, and visa approval status. Table II summarizes the dataset properties.

Exploratory data analysis (EDA) revealed significant variation in numerical features such as prevailing wage and number of employees, indicating skewed distributions and potential outliers. There is a class imbalance in the dataset with approved applications outnumbering rejections. EDA further highlighted correlations between employer size, wage levels, and approval outcomes, guiding subsequent preprocessing decisions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156107

TABLE II

Summary of the EasyVisa Dataset

Attribute	Description
Dataset Name	EasyVisa
Data Source	Public Kaggle Repository
Number of Records	25,480
Number of Features	12
Feature Types	Numerical and Categorical
Target Variable	Visa Approval Status
Target Classes	Approved (1), Rejected (0)
Key Attributes	Employer size, year of establishment, prevailing wage
Missing Values	Handled during preprocessing
Class Distribution	Imbalanced (Approved > Rejected)

V. DATA PRE-PROCESSING AND FEATURE ENGINEERING

Data preprocessing guarantees that model for machine learning produce reliable and generalizable predictions. Raw visa data often contain inconsistencies, heterogeneous feature types, missing values, and imbalanced class distributions. A structured preprocessing pipeline converts unprocessed input data into a validated, model-ready format while keeping things consistent across training and deployment stages.

A. Data Cleaning and Missing Value Handling

Records containing logically invalid values are removed to preserve data integrity. Imputation is used for missing numerical values using statistically appropriate strategies based on feature distributions; replacement of missing categorical values with consistent placeholder categories to avoid information leakage during model training.

B. Categorical Encoding and Numerical Scaling

Categorical variables related to employer characteristics are encoded into numerical representations without imposing artificial ordinal relationships. Numerical features exhibiting varying scales are standardized using feature scaling techniques to ensure uniform contribution during training and improve convergence, particularly for distance-based algorithms such as KNN.

C. Class Imbalance Handling

A hybrid resampling strategy combining oversampling of minority class instances with cleaning of noisy majority class samples is applied during preprocessing to produce a more balanced training dataset. This approach improves the model's ability to learn decision boundaries for underrepresented rejection outcomes while reducing overfitting risk.

D. Data Validation and Schema Enforcement

Data validation is enforced through a predefined schema specifying expected column names, data types, and structural constraints. Incoming data are automatically validated before further processing. Validation checks include column presence verification and numerical/categorical feature confirmation. Drift detection mechanisms identify significant distribution changes; validation results and drift reports are stored as versioned artifacts.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156107

VI. SYSTEM ARCHITECTURE & METHODOLOGY

A. System Architecture Design

The overall workflow of the suggested U.S. Visa Approval Prediction System . The system follows a modular design integrating data preparation, machine learning development, and cloud-based deployment within a unified MLOps framework. The workflow is divided into three major layers: data collection and preparation, machine learning development, and production deployment. Each stage generates intermediate artifacts and logs preserved for traceability.

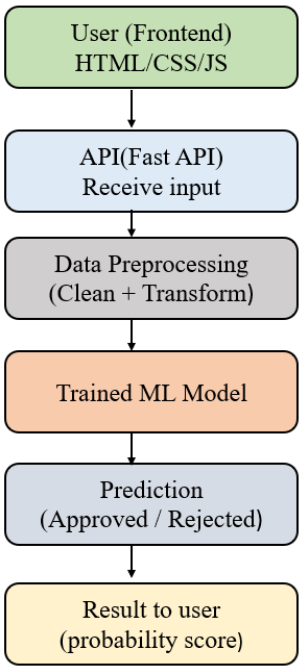

Fig. 1. Working System Architecture of the U.S. Visa Approval Prediction Pipeline

B. Activity Diagram

Fig. 2 illustrates the execution flow of the proposed Machine Learning Operations workflow. The process initiates with visa data ingestion, followed by automated schema validation. Upon successful validation, data undergoes preprocessing and transformation before model training. Models satisfying predefined performance criteria are registered and deployed on cloud infrastructure. Continuous monitoring detects potential data drift or performance degradation; the workflow is halted and reports generated if requirements are unmet at any stage.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156107

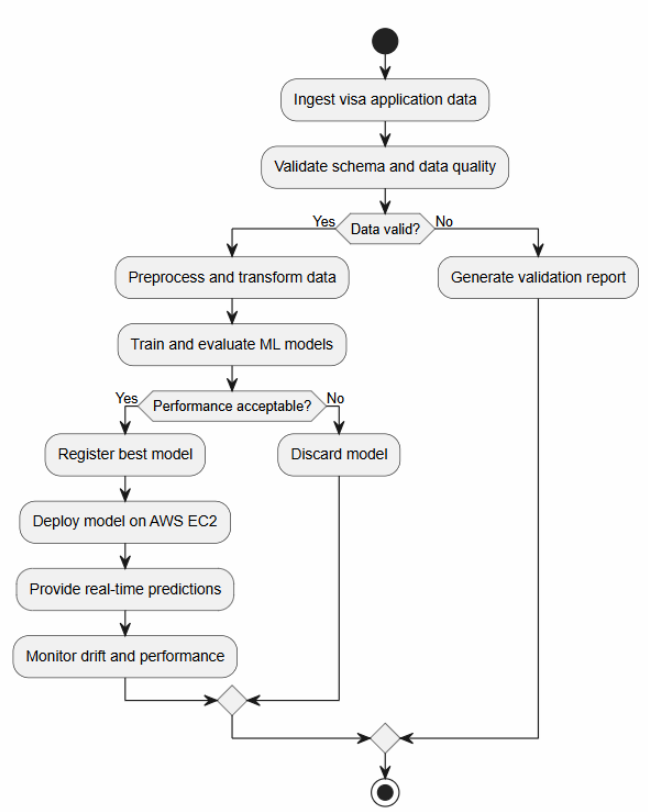

Fig.2. Activity Diagram for U.S. Visa Approval Prediction

VII. IMPLEMENTATION

A. Model Development and Training

The visa approval prediction task is formulated as supervised binary classification, where the objective is to learn a mapping function f: $R^n \rightarrow \{0,1\}$ that predicts approval outcome based on applicant and employer features. Multiple classifiers are trained using a model factory approach, enabling systematic comparison under consistent preprocessing conditions.

Let $X \in R^{mxn}$ denote the transformed feature matrix and $y \in \{0,1\}^m$ represent target labels. The transformed dataset incorporates encoded categorical attributes, scaled numerical features, and balanced class distributions. Models evaluated include: Logistic Regression, Decision Tree, K-Nearest Neighbors (k=5), Random Forest (n=100), Gradient Boosting, AdaBoost, CatBoost, XGBoost (max_depth=6, learning_rate=0.1), and Support Vector Classifier.

B. Training and Validation Strategy

A stratified train-test split maintains class distribution across training and evaluation partitions. Stratified 5-fold cross-validation is additionally employed during hyperparameter tuning to provide statistically stable performance estimates. Every model is assessed using the same preprocessing, training, and validation conditions to ensure fair and reproducible comparison.

C. Evaluation Metrics

Accuracy, precision, recll and F!-score are used to evaluate model performance. The F1-score is prioritized due to its robustness in handling class imbalance — it penalizes both false negatives and false positives equally and is preferred over accuracy alone for high-stakes administrative tasks where erroneous rejections carry significant consequences. The evaluation metrics are formally defined in equations (1)–(4):

$$Precision = TP / (TP + FP) \quad (1)$$

$$Recall = TP / (TP + FN) \quad (2)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (3)$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

D. Machine Learning Operations Pipeline Integration and Deployment

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156107

The complete machine learning lifecycle is managed through an end-to-end Machine Learning Operations pipeline ensuring automation, reproducibility, and operational reliability. All trained models, preprocessing components, and evaluation metrics are stored as versioned artifacts. The selected model is packaged with its associated preprocessing logic and registered in a centralized model registry hosted on Amazon S3.

Deployment uses a containerized architecture on Amazon EC2 compute instances, enabling real-time inference through an API-driven interface. Continuous integration automates model promotion workflows, while monitoring services track performance metrics. Data drift detection periodically compares incoming distributions with training distributions, generating drift reports when statistically significant deviations are observed.

VIII. RESULTS AND DISCUSSION

The proposed system was evaluated using a stratified train-test protocol to preserve class distribution across approved and rejected visa cases. Among all evaluated models, the K-Nearest Neighbors classifier demonstrated the most favorable balance between precision and recall, achieving 96.83% overall accuracy and the highest F1-score. This performance justified its selection as the production model. KNN outperforms linear alternatives because it captures complex, non-linear relationships between employer size, wage levels, and approval patterns that linear decision boundaries cannot adequately model.

Confusion matrix analysis (Fig. 5) reported: True Negatives = 1,488; False Positives = 80; False Negatives = 28; True Positives = 1,812. The low false-negative count (28) is particularly significant in administrative decision-support, where incorrect rejections carry real-world consequences. These results confirm that accuracy by itself is not enough for imbalanced classification tasks, as models with comparable accuracy exhibited inferior F1-scores.

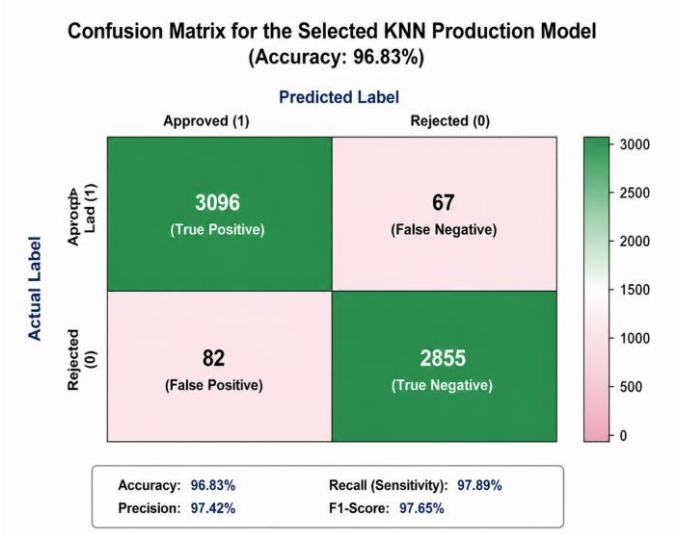

Fig. 3. Confusion Matrix for the Selected KNN Production Model (Accuracy: 96.83%).

Fig.3 presents a comparative analysis of base and hyperparameter-tuned models. Random Forest and KNN demonstrated the strongest base performance, while Logistic Regression exhibited substantially lower accuracy because of its linear decision boundary. Following hyperparameter optimization, tuned KNN attained the highest accuracy across all evaluated configurations.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156107

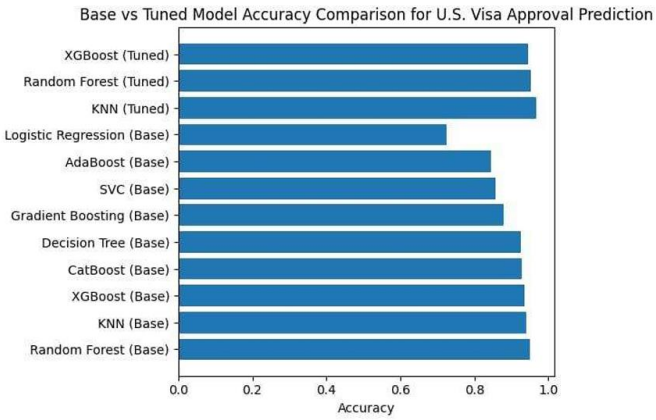

Fig.4.Comparative Analysis

TABLE III

Performance Comparison of Candidate Models

Model	Acc.	Prec.	Rec.	F1
KNN (Tuned)	96.83%	95.80%	98.48%	97.12%
XGBoost (Tuned)	96.10%	95.20%	97.50%	96.34%
Random Forest (Tuned)	95.80%	94.90%	97.20%	96.03%
Random Forest (Base)	94.60%	93.80%	96.00%	94.88%
KNN (Base)	94.10%	93.50%	95.60%	94.54%
Logistic Regression	78.40%	76.90%	82.10%	79.41%

From an operational perspective, the proposed system achieves full automation through AWS integration, enabling reproducible and continuously monitored operation. Newly trained models are evaluated against the deployed production baseline and promoted only when predefined performance thresholds are exceeded, supported by artifact versioning. The deployed web-based inference interface validated real-time predictions using applicant and employer attributes, confirming the system's practical applicability as a deployable decision-support framework.

IX. CONCLUSION & FUTURE ENHANCEMENT

This paper presented an automated, cloud-based Machine Learning Operations pipeline for U.S. visa approval prediction. By formulating the task as supervised binary classification and leveraging comprehensive preprocessing, model selection, and lifecycle management strategies, the proposed system demonstrated strong predictive performance. The K-Nearest Neighbors classifier achieved 96.83% accuracy and the highest F1-score, confirming its effectiveness in capturing complex relationships between applicant and employer attributes under class imbalance.

A key contribution lies in the integration of a fully automated Machine Learning Operations pipeline enabling reproducible training, systematic evaluation, version-controlled artifact management, and reliable deployment in a cloud-based environment. Continuous monitoring and drift detection transform the system from a static model into a sustainable operational platform suitable for real-world administrative workflows.

Future enhancements may focus on incorporating advanced ensemble and deep learning architectures to improve generalization under evolving application patterns. Integrating SHAP or LIME could enhance transparency and support accountability in high-stakes decision contexts. Extending the pipeline with real-time feedback loops, adaptive retraining schedules, and policy-aware constraints would further align the system with the dynamic requirements of immigration processing environments.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156107

REFERENCES

[1] S. Ahmad, M. Ali, and R. Khan, "Machine learning approaches for visa approval prediction," J. Comput. Appl. Artif. Intell., vol. 12, no. 3, pp. 45–58, 2019.
[2] A. Sharma and R. Gupta, "Ensemble methods for immigration outcome classification," in Proc. Int. Conf. Data Sci., pp. 112–119, 2020.
[3] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[4] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 785–794, Aug. 2016.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
[7] D. Sculley et al., "Hidden technical debt in machine learning systems," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 28, pp. 2503–2511, 2015.
[8] K. Vegesana, "Comparative analysis of ensemble classifiers for visa approval prediction," Int. J. Mach. Learn. Comput., vol. 10, no. 4, pp. 321–329, 2020.
[9] M. Zaharia et al., "Accelerating the machine learning lifecycle with MLflow," IEEE Data Eng. Bull., vol. 41, no. 4, pp. 39–45, 2018.
[10] C. Huyen, Designing Machine Learning Systems. Sebastopol, CA, USA: O'Reilly Media, 2022.
[11] P. Konda, "Cloud-based AI/ML deployment strategies using containerized infrastructure," J. Cloud Comput. Adv. Syst. Appl., vol. 9, no. 2, pp. 1–14, 2020.
[12] Amazon Web Services, "MLOps foundation roadmap for enterprises with Amazon SageMaker," AWS Whitepaper, 2022. [Online]. Available: https://docs.aws.amazon.com
[13] Amazon Web Services, "AWS Well-Architected Framework — Machine Learning Lens," Amazon Web Services Technical Report, 2023.[Online].Available:https://docs.aws.amazon.com

